EUDAT Conference Posters

FINAL PROGRAMME | REGISTRATION | POSTERS & DEMOS | VENUE & ACCOMMODATION

1 | eInfraCentral – Helping users discover and access Europe's e-infrastructure services

Author: Orsolya Gulyas

Affiliation: European Future Innovation System (EFIS) Centre

Abstract:
eInfraCentral (EIC) is a coordination and support action funded from the European Union's Horizon 2020 research and innovation programme. Its mission is to ensure that by 2020 a broader and more varied set of users discovers and accesses the existing and developing e-infrastructure capacity. The underlying idea is to create a marketplace.
The project has three-fold objections. First, to structure an open and guided discussion between e-infrastructures to consensually define a common service catalogue. Second, to develop a one stop shop/single-entry point portal for users to browse the service catalogue and enhance the monitoring of key performance indicators that focus on availability and quality of services and user satisfaction. Third, to draw policy lessons for a future European e-infrastructure market place as an extension of the common service catalogue and portal (incorporating a larger number of e-infrastructures).
The poster to be presented during the DI4R is targeted at three audiences: the e-infrastructure services providers (pan-European, regional and national, monothematic or poly-thematic, uni-disciplinary or multi-disciplinary, etc.), b) virtual research environments (VREs), and c) potential users of the e-infrastructure services.
The poster aims to answers to some potential questions from the users:
"How eInfraCentral project will respond to the need of the European researchers on digital services?"
"What are we developing in the context of the European Open Science?"
"What will be offered?"

2 | Data Curation and Provenance for the EUDAT Collaborative Data Infrastructure

Author: Alexander Atamas

Affiliation: Data Archiving and Networked Services (DANS), The Hague, Netherlands

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Data curation plays a significant role in research data management and includes selection, preservation, maintenance, collection and archiving of digital data. Practically, data curation involves establishing and developing long-term storage repositories of digital assets for processing by scientists and other communities. Therefore, data curation within the EUDAT Collaborative Data Infrastructure (CDI) is of great interest for the research community. The presented work describes policies for digital data management as prospective candidates of machine-executable services for EUDAT B2SAFE platform. Data curation policies for two use cases have been developed and mapped to requirements of HERBADROP and GEOFON research projects. It has been demonstrated that the developed policies are strong candidates for implementation in the two use cases. Furthermore, a gap analysis has been carried out against the SCAPE catalogues of policies.
Provenance of research data is important for tracking origins, ownerships and modifications of data over the data lifecycle. The concept of provenance guarantees that the data sets made available for sharing and exchange are reliable and hence all data transformations and results obtained using the data sets could be reproduced. A HTTP template-based service collecting provenance is developed in our work package to be used by any service from the EUDAT CDI portfolio. The service defines an API for the clients to generate provenance data based on particular templates (Notation3 format) based on the PROV Ontology (PROV-O), which are made available by its operator. The gathered provenance data then can be queried by any interested EUDAT service by using HTTP protocol. The provenance API is demonstrated by applying it to two use cases: the process of persistent identifier (PID) generation in the B2SAFE service; the workflow of the B2SAFE replication.

Nowadays, research data often undergo rather intensive lifecycle. Individual data sets for one reason or another are modified, updated or recalculated. Hence, in the scientific community, there is a need for versioning functionality for proper data curation and provenance of research data. Research data sets can constitute a wide spectrum of different formats, for example, papers in PDF format, source codes, log files, images, and videos. Software version control systems like “git” are not able to deal efficiently with large size binary data sets. Therefore, in this work a versioning functionality has been designed and implemented on B2SAFE service. PIDs are employed to identify and access versions independently of physical storage location. In our implementation, a new version is made cross-linked with previous version for the sake of easy navigation between versions.

Having well-designed and clearly expressed data policies including tools for the automated policies execution, as well as having provenance records of data acquisition, transformation and distribution is important for the promotion of data infrastructure platforms such as B2SAFE to the state of purposeful, trusted and well-managed IT services. The EUDAT Collaborative Data Infrastructure is expected to adopt approaches, techniques and tools for data curation developed by EUDAT projects, and to apply them for new research user communities in addition to those involved in EUDAT pilots.

3 | ENES Climate Analytics Service (ECAS), A Contribution to the EOSC-HUB

Author: Sofiane Bendoukha
Affiliation: Deutsches Klimarechenzentrum (DKRZ)

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Within the EOSC-HUB project, an integrated catalogue of services, software and data from the EGI federation, EUDAT CDI, INDIGO-DataCloud and major research e-Infrastructures will be delivered. The ENES Climate Analytics Service (ECAS) is a thematic service, that will take part of the catalogue that constitutes the HUB of the future European Open Science Cloud.

The main objective of ECAS is to allow scientific end-users mainly from the climate data community to perform data analysis experiments on large volumes of climate data.
ECAS follows a Persistent Identifier (PID)-enabled, server-side and parallel approach. Based on experiences within the climate data community, ECAS will open up processing capabilities also for use by other disciplines.

ECAS relies on different service and software components that proved their efficiency in terms of data management and processing.
The essential components of the ECAS service are: B2HANDLE and Ophidia framework. On the one hand, Ophidia is a mature, complete and stable service for data analytics, as a result of a long internal validation phase with end-users at Fondazione CMCC. On the other hand, B2HANDLE is a distributed service, offered by EUDAT and designed to contribute to data persistency by maintaining opaque, globally unique Persistent Identifiers (PIDs).
The integration between Ophidia and B2HANDLE plays an important role within the ECAS ecosystem. It will enable basic data provenance tracking by establishing PID support through the whole chain, and thereby improving reusability, traceability, and reproducibility.

Besides B2HANDLE and Ophidia, ECAS will also rely on other services from EUDAT service suite:

- B2DROP to synchronize data outputs between peers
- B2ACCESS to authenticate and authorize users
- B2SHARE to store and share small data sets

4 | Welcome DuraCloud Europe: Secure Storage in the Cloud

Author: Andrea Bollini [1], Erin Tripp [2], Heather Greer Klein [2], Susanna Mornati [1], Claudio Cortese [1]
Affiliation: 1: 4Science Srl, 2: DuraSpace

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
DuraCloud (http://www.duracloud.org/) is open technology developed by DuraSpace, released [1, 2] in 2010 as open source under the Apache 2.0 license, that makes it easy for organizations and end users to use content preservation services in the cloud. DuraCloud leverages existing cloud infrastructure to enable durability and access to digital content. It is particularly focused on providing preservation support for academic libraries, academic research centers, and other cultural heritage organizations.

Digital preservation is a complex ecosystem. DuraCloud can serve as an important component of this ecosystem, supporting funder policy compliance and enabling best practices in data management to support reuse and data integrity, especially for research data.

The service builds on expert cloud storage providers by overlaying the access functionality and preservation support tools that are essential to ensuring long-term access and durability. DuraCloud offers cloud storage across multiple providers and geographic regions, and offers compute services that are key to unlocking the value of digital content stored in the cloud. DuraCloud allows users to implement preservation policies and data curation strategies for research data providing services that enable digital preservation, data access, transformation, and data sharing.

Since 2011 the scholarly community has benefited from the DuraSpace managed DuraCloud service in the United States. To allow a broader world-wide access to these services with close time-zone support and storage proximity, in November 2017 4Science became the first member of the Certified DuraSpace Partner Program, delivering DuraCloud services [3]. Now, from the start of 2018, 4Science provides a managed DuraCloud Service in Europe allowing preservation and content storage service, complying with the European Commission General Data Protection Regulation (GDPR) and offering application support in additional time zones and languages. The DuraCloud service by 4Science provides affordable, transparent, vendor independent, and cost-predictable access to the cloud storage facilities in Europe making it easier for institutions to take advantage of the infinite scalability offered by the cloud.

[1] DuraCloud Now Available as Open Source: Preservation Support and Access Services Built on Cloud Infrastructure - http://www.duraspace.org/articles/1677
[2] DuraCloud GitHub Repositories - https://github.com/duracloud
[3] 4Science Offering DuraCloud Services in Europe as Certified DuraSpace Partner - http://duraspace.org/taxonomy/term/163

5 | The Europeana Data Pilot: outcomes and conclusions

Author: Nuno Freire
Affiliation: Europeana / INESC-ID

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Europeana is Europe's digital platform for cultural heritage, providing access to to over 54 million of digitised cultural resources from a wide range of cultural institutions from over 3700 cultural heritage institutions, ranging from books, photos and paintings to television broadcasts and 3D objects. It seeks to enable users to search and access knowledge in all the languages of Europe. This is done either directly, via its web portals, or indirectly, via third-party applications built on top of its data services (search APIs and Linked Open Data).

One of Europeana's most recent lines of action is to facilitate the research on the aggregated data resources. This work is conducted in the scope of Europeana Research, where issues affecting the research re-use of cultural heritage data and content (such as licensing, interoperability and access) are addressed.

Europeana is currently interested in investigating if, and how, research data e-infrastructures can support its mission to address the requirements for research use of its data resources. The vision is that by leveraging on other European level e-infrastructures for research data, it will be able to reach all potentially interested researchers from all scientific disciplines. Without generic and cross discipline data e-infrastructures, such as EUDAT, Europeana would have to work with several other e-infrastructures or provide its own research data infrastructure.

With this objective in mind, Europeana has been conducting a data pilot with EUDAT. In the pilot, Europeana is using, as case study, the Europeana Newspapers corpus - one of its datasets that has attracted the most interest from researchers. The pilot investigates the use of the EUDAT services for sharing the corpus through the right mechanisms for its use in research.

The poster presents the results obtained after conducting two iterations of a prototype for publication of cultural heritage datasets from Europeana into EUDAT.

6 | Bringing Europeana and CLARIN together: Dissemination and exploitation of cultural heritage data in a research infrastructure

Author: Nuno Freire
Affiliation: Europeana / INESC-ID

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
We present the joint work by Europeana (http://www.europeana.eu), a European cultural heritage (CH) infrastructure, with CLARIN (www.clarin.eu), a European research infrastructure, to make promptly available for research use the vast data resources that Europeana has aggregated in the past years.

Europeana provides access to digitised cultural resources from a wide range of institutions all across Europe. It seeks to enable users to search and access knowledge in all the languages of Europe, either directly via its web portals, or indirectly via third-party applications leveraging its data services. The Europeana service is based on the aggregation and exploitation of (meta)data about digitised objects from very different contexts. The Europeana Network has defined the Europeana Data Model (EDM) to be used as its model for interoperability of metadata, in line with the vision of linked open vocabularies. One of the lines of action of Europeana, is to facilitate research on the digitised content of Europe's galleries, libraries, archives and museums, with a particular emphasis on digital humanities.

CLARIN (Common Language Resources and Technology Infrastructure) is a networked federation of language data repositories, service centres and centres of expertise.
CLARIN aggregates metadata from resource providers (CLARIN centres and selected "external" parties), and makes the underlying resources discoverable through the Virtual Language Observatory (VLO) to provide a uniform experience and consistent workflow.
The VLO can also serve as a springboard to carry out natural language processing tasks via the Language Resource Switchboard (LRS), allowing researchers to invoke tools with the selected resources directly from its user interface. The potential inclusion of many new CH resources by 'harvesting' metadata from Europeana, opens up new applications for CLARIN's processing tools.

CLARIN and Europeana do not share a common metadata model, and therefore a semantic and structural mapping had to be defined, and a conversion implemented on basis of this. CLARIN's ingestion pipeline was then extended to retrieve a set of selected collections from Europeana and apply this conversion in the process. Several infrastructure components had to be adapted to accommodate the significant increase in the amount of data to be handled and stored. Currently about 775 thousand Europeana records can be found in the VLO, with several times more records expected in the foreseeable future. Currently, about 10 thousand are technically suitable for processing via the LRS. Relatively straightforward improvements to the metadata on the side of Europeana and/or its data providers could substantially increase this number. CLARIN is working with Europeana to implement such improvements. More tools are also expected to be connected to the LRS in the short to mid-term, which is also expected to lead to an increased 'coverage'.

As a next step, CLARIN can extend and refine the selection of included resources, and Europeana can adapt their data and metadata to optimally serve the research community. CLARIN's experience and potentially part of its implementation work can be applied to integrate Europeana with other resource infrastructures.

7 | Exploring the costs and scalability of research data management services

Author: Claudia Engelhardt (1), Sven Bingert (2), Harald Kusch (3), Stefanie Wache (4)

Affiliation: (1) Göttingen State and University Library (SUB), (2) Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG), (3) University Medical Center Göttingen (UMG), (4) University Medical Center Göttingen (UMG)

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Claudia Engelhardt, Göttingen State and University Library (SUB), claudia.engelhardt@sub.uni-goettingen.de
Sven Bingert, Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG), sven.bingert@gwdg.de
Harald Kusch, University Medical Center Göttingen (UMG), harald.kusch@med.uni-goettingen.de
Stefanie Wache, University Medical Center Göttingen (UMG), stefanie.wache@med.uni-goettingen.de

How can we support the research data management at a diverse research campus in a sustainable manner … and what will it cost? Two crucial aspects that need to be considered in order to solve this issue are the scalability or generalisability of research data services and a cost model to determine the respective efforts. The two-year project Göttingen Research Data Exploratory (GRAcE), funded by the German Federal Ministry of Education and Research, examines these topics and develops a planning instrument to support the campus-wide establishment and sustainable operation of research data management services. This toolset is developed based on the Göttingen Campus as an example, but is intended to also be applicable to other campuses.
In order to determine the costs of data management, GRAcE dissects the different stages of the research data lifecycle and the relevant tasks in each of them. By that the efforts – in terms of technology as well as staff – can be estimated for every task and then aggregated first on the level of the project, then the institute and so forth. A challenge in this regard is the diversity of research methods between different and often also within fields of study. Depending on the research question and the approach taken to investigate it, but also on the skills and experience of the researchers and other staff involved, the effort for data management can vary considerably. Another challenge lies in the fact that, although a number of roles (like data manager, data steward, data scientist etc.) covering different skills and competence sets in the field are beginning to emerge, data management tasks often are fulfilled by staff members who do not have an explicit data management role such as PhD students or IT technicians – so that the respective efforts are hidden. GRAcE aims to make these visible and provide a realistic cost estimation. The work in this area is led by the Department of Medical Informatics of the University Medical Center Göttingen (UMG).
Regarding the scalability, GRAcE focuses on the generalisability of data management services developed in and for a specific research context. Large joint research projects such as the DFG funded Collaborative Research Centres (CRC), for example, often implement subject-specific RDM solutions. GRAcE aims to define criteria that help to determine if and if yes, to which degree such tailor-made solutions can be adapted for a wider use and transferred into a more generic context. Subsequently, the relevant parameters for the adaption for, transfer to and integration into a research data management infrastructure covering the multi-disciplinary range of subjects on the campus will be examined. The research in this field is conducted by the State and University Library Göttingen (SUB) and the Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG).
The project closely collaborates with the Göttingen eResearch Alliance (eRA), an initiative that provides eResearch and data management support for Göttingen campus. The results will contribute to adaptions of existing and the development of new services provided by the eRA.

8 | OpenAIRE: supporting the EC's Open Data Pilot in H2020

Author: Pedro Principe
Affiliation: UMINHO

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
OpenAIRE's ambitious plan is to foster the social and technical links in scholarly communication to make science open and reproducible a reality in Europe and beyond.
In fact, OpenAIRE addresses key aspects and challenges of the currently transforming scholarly communication landscape and actively seeks and promotes new solutions that better suit the needs of researchers, innovators, the public and funding bodies, relevant to new technologies and expanding amounts of information.
OpenAIRE supports the implementation of the EC's requirements and recommendations on Open Access and Open Research Data through its network of 33 National Open Access Desks, providing effective outreach and engagement to different stakeholders at the national and institutional levels, and assistance in fostering collaboration between different disciplines.
OpenAIRE2020, the third phase of the OpenAIRE project series, focuses on research data management, data archiving and sharing practices and supports the implementation of the Open Data Pilot in Horizon2020, which aims to improve and maximize access to and re-use of research data generated by EU-funded projects. In particular, through its locally embedded and international connected network of NOADs, OpenAIRE is capable of rolling out a supporting network to facilitate the Open Data Pilot across Europe. It developed a set of useful resources on policy matters, the implementation of the Pilot, tools and information to European researchers, project managers and research supporting staff.
As such OpenAIRE accommodates the translation and implementation of the EC OS and ORD policy to the adaptation of data management practices across Europe, streamlining implementation at national level and as such employs a node of expertise in each country.
In Summer 2017, OpenAIRE ran a survey on the European Commission's approach to Data Management Plans to evaluate and give feedback on its current template. The survey dealt with topics such as:
• Developing and reviewing Data Management Plans
• FAIR principles for research data
• discipline-specific guidelines
• common standards for DMPs
Among the outcomes of this survey emerges the need for internal training courses for supporting staff to advise beneficiaries and review DMPs, and also for guidance on potential costs of RDM. Using networks such as the OpenAIRE NOADs and RDA-Europe will, therefore, be key to disseminate up-to-date information on the EC's requirements to the wider RDM support community as well as providing services acting as a pan-European coordinator and aggregator node, reaching out to national nodes and coordinating them on open data implementation. It highlights again the fact that policy should always be backed-up with acute infrastructure and services. By delivering a cultural and technological shift towards common Open Science practices, OpenAIRE aligns the implementation of Open Science policies as well as customized solutions and unified services providing a flexible framework that can adapt to the wide variety of challenges presented.

9 | The EOSC as a knowledge marketplace: the example of ISIDORE

Author: Dumouchel Suzanne, Ribbe Paulin
Affiliation: Huma-Num (CNRS)

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
We would like to draw the benefits of ISIDORE integration into the EOSC which could be, as a suggestion, a knowledge marketplace.
ISIDORE, developed by TGIR Huma-Num (CNRS) in France, impulses a virtuous research circle for SSH researchers. This service collects, enriches and provides unified access to digital documents and data from the humanities and social sciences in the whole Europe.
ISIDORE harvests structured and unstructured data: bibliographical records, metadata, integral text from digital publications, corpus, databases and scientifical news accessible on the web.
Once harvested, those informations are enriched and standardized in different languages, by crossing with referentials (vocabulary lists, thesaurus) produced either by the scientific community, either by research institutions. Those enrichments allow to link the data between each other.
Launched in 2009, more than 5800 sources are already harvested, for a total of more than 5 millions documents
The re-exposure of the enriched metadata follows, in turn, the principles of Web of data (RDF) claimed by the movement of provision of public data as data.gov (USA) and data.gov.uk (UK). Thanks to this feature, ISIDORE is different from a simple search engine : it offers to the whole community to enrich constantly its own data.
Otherwise, SSH publications are numerous: researchers don't have main revues as in other fields and it exists several small revues. That means, it is very difficult for researchers to find information and publications or to make them more visible. ISIDORE is the only tool in Europe able to crawl all the sources. So it is very important to scale it to European researchers, by integrating it in the EOSC.
We consider the EOSC as a knowledge marketplace, which will share and spread tools and good practices from the whole European Research Area. It needs common policy, interoperability, easy access. In this perspective, the marketplace could contain a public part open for every researcher and public institution and a market perspective to let companies use our tools and data. By doing that, sustainability of the EOSC is ensured. We envision three levels:
_ a storage one (interoperability of computing centres)
_ a services one (the marketplace)
_ a communication and training one (an appstore). This third aspect builds the "knowledge marketplace": the ability to use research data, research tools, so to develop and increase knowledge, but also the ability to share knowledge, to comment those tools, to create some communities around each (or more) ones. Bridges between researchers and citizens which strenghten the notion of "knowledge" is very important.
The EOSC offers digital solutions by containing a collection of software, tools, services, datasets, publication repositories and learning & training material and will establish visibility for them. Our conception of EOSC is highly linked to the needs of the SSH researchers to be able both to find resources and make them visible and to establish links with the civil society, and ISIDORE answers to the same objectives.

10 | CETAF stable identifiers for specimens

Author: Anton Güntsch
Affiliation: Freie Universität Berlin, Botanischer Garten und Botanisches Museum Berlin

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Natural history collections are estimated to contain more than 2.5 billion specimens worldwide and institutions throughout Europe hold the major proportion of this priceless heritage. Being a tremendously important basis for research and the only physical evidence of the past occurrence of organisms in space and time, biological collection objects need to be consistently referenced with globally unique and stable identifiers.
The Information Science and Technology Committee (ISTC) of the Consortium of European Taxonomic Facilities (CETAF, http://www.cetaf.org) has defined a simple and future-oriented identifier system for specimens based on HTTP-URIs and Linked Data principles, building a bridge to rapidly developing semantic web technologies.
Each individual collection object as well as its associated information resources (e.g. multimedia, RDF, webpages) is designated by a URI chosen and maintained by the institution owning the specimen. Identifiers are typically composed of an intuitions' web domain, a meaningful subdomain, a path to classes of similar objects, and local objects identifiers (e.g. the object barcode). Since physical objects cannot be transferred via the Internet, users trying to access an object using a web-browser will be redirected to a human-readable representation of the object, typically an html web-page. Likewise, software-systems requiring machine-processable representations will be redirected to an RDF-encoded metadata record.
To date, over 20 Million specimens from 13 European collections are accessible via the CETAF stable identifier system. For collections wanting to implement stable identifiers, the initiatives offers best practices documents, a web-services for testing URIs, and software for implementing local redirection mechanisms.

11 | Creating simple data management plans

Author: Adil Hasan

Affiliation: UNINETT Sigma2

Abstract:
The process of creating a data management plan requires understanding the requirements outlined in often brief guidelines. Within EUDAT, and in collaboration with openAIRE we are developing a tool that will allow researchers to create plans in an easy manner. The tool will make use of some of the EUDAT services such as the Data Project Management Tool and B2Access to provide better integration with EUDAT data management services. The poster will describe the tool, its integration and the workflow as well as some model use-cases. We will provide a link to allow interested users to try out the tool and provide feedback.

12 | EUDAT Data Subscription Service in the EuroArgo case

Author: Jani Heikkinen

Affiliation: CSC - IT Center for Science Ltd

Abstract:
The poster/demo describes the EuroArgo case in which marine domain researchers are interested in changes in accumulating datasets due to the effects of the changes in dataset query results. Furthermore, the poster/demo shows how EUDAT Data Subscription Service can be applied in this use case to help researchers focus on other tasks at hand instead of waiting and monitoring a change to happen. Through the EUDAT Data Subscription Service, an application or a community portal serving the end-user can refer to a data query which is automatically processed whenever new data is added to the data origin repository. The user is notified when new data matching her subscription is found. In addition, the poster/demo describes an integration across research- and e-infrastructures.

13 | The WSL Environmental Data Portal EnviDat in the context of Pan-European Research Data Management Services

Author: Ionut Iosifescu

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Affiliation: Swiss Federal Research Institute for Forest, Snow and Landscape Research WSL

Abstract:
The amount and quality of environmental data is rapidly increasing worldwide. The Swiss Federal Research Institute WSL has a long tradition in data collection, in particular in the areas of forest ecosystems, snow and natural hazards research. The data sets collected by WSL researchers include time series and spatial samplings spanning over 100 years. WSL operates a comprehensive network for environmental research that includes more than six thousand observation sites for studying the terrestrial environment, including the terrestrial carbon cycle and its changes in a changing climate. Such long-term environmental monitoring datasets are particularly valuable towards obtaining an integrated view of the Earth System, while data sharing encourages new national, European and international collaborations. Unfortunately however, a unified portal allowing simple and efficient access to all kinds of environmental data sets and their metadata does not currently exists. In Switzerland, for example, centralized access to environmental data sets is still missing despite plans to largely move towards open data in the mid- to long-term.
To this end, WSL is developing EnviDat, an overarching data portal for facilitating the management, search and user-friendly access to its rich reservoir of environmental data. The portal's main functional requirements include data discovery through metadata and map search, publishing of datasets with Digital Object Identifiers (DOI) and provision of a repository for diverse data types, whereas data curation and quality control remain with the experts.
The EnviDat core design principles are focusing on usability and user-friendliness. Moreover, additional relevant non-functional requirements such as reliability, security and maintainability are considered in its multi-server system architecture. The EnviDat conceptual framework, presented in the poster, not only refines existing functionalities, but also tightly integrates several principles such as the connection to the wider research data management community and, where possible, the adoption of best practices and standards in data sharing. High importance is therefore laid on future technical interoperability with the wider pan-European data management community. Basic interoperability can be achieved by leveraging well-known community software, as for instance CKAN, whereas long-term interoperability could be achieved by incorporating key European data management services, such as B2FIND, B2DROP and B2SHARE.
Existing institutional thematic repositories such as EnviDat may therefore help increase pan-European visibility of environmental monitoring and research data through metadata harvesting and long-term community storage options for highly curated datasets. However, there remain substantial challenges resulting from aligning an institutional repository with an overarching EU data framework at both technical and strategic levels, as for example related to the preview and visualization of long-term monitoring data. EnviDat welcomes these challenges and aims at facilitating the sharing of environmental data with the pan-European community.

14 | TSD and the Clinical Trial Pilot in EUDAT

Author: Maria Francesca Iozzi
Affiliation: UNINETT/SIGMA2

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Clinical trials involve human volunteers and patients and are regarded as the most effective way to add knowledge to the growth of the medical domain. According to a study protocol, participants receive specific interventions that may be medical products, such as drugs or devices. International European clinical trials are supported by ECRIN (The European Clinical Research Infrastructures Network) a distributed research infrastructure established to help researchers with high quality clinical research. ECRIN offers support to multinational clinical research projects through information and consultancy, and by providing a variety of support service.
As member of the CORBEL project (www.corbel-project.eu/) ECRIN is also investigating solution to store and share individual-level patient clinical trials data (IPD). Clinical study data are stored as pseudonymised, coded data sets with additional documents that contain metadata and explanations necessary for analysis of the data. Clinical trials data is sensitive data and therefore, they have to be stored and processed under consideration of data protection rules (e.g. informed consent, restricted access), to prevent any re-identification of study subjects. The aim of our Clinical Trial Pilot is to evaluate EUDAT services (B2SAFE for long-term storage and B2SHARE for sharing anonymised data) that were created for open data exchange for the compliant usage for sensitive data from clinical trials.
The General Data Protection Regulation (GDPR) is now setting the stage for accessing, sharing and processing of sensitive personal data in Europe. Several strategies have been adopted locally to comply with national privacy regulations. But there is still the need to implement policies and the corresponding technologies that effectively allow cross border, inter-disciplinary research on personal sensitive data. In recent years, EUDAT has investigated the possibility for using state-of-art compliant solutions, currently offered on national settings, for wider use across european-wide communities. For the Clinical Trial Pilot, the candidate service to store and analyse these data in a secure compliant infrastructure is the TSD - Service for Sensitive Data, developed and operated by the University of Oslo and Sigma2 (Norway). TSD is a remote access solution to allow single users and institutions to store, analyse, compute sensitive research data. All the data are stored in a centralised storage solution secured by a strong firewall. The Clinical Trial Pilot is currently testing the TSD solution to evaluate if it can be used to analyse clinical trials data (by using R software inside the TSD) without risking the re-identification of the data subjects. Simultaneously EUDAT is investigating how to offer TSD to the research communities dealing with sensitive data (e.g. ECRIN for clinical data) through proper mechanism of identity management and through a proper legal framework compliant with the European data protection regulation (e.g. GDPR) and GCP (Good Clinical Practice).

15 | Sensitive personal data: activities in the EUDAT CDI and future scenarios

Author: Maria Francesca Iozzi

Affiliation: UNINETT/SIGMA2

Abstract:
In recent years EUDAT has collected user requirements through pilot calls, and some of them were concerning the management of sensitive personal data. Indeed the use of human data for research is increasing in several disciplines, thus requiring advanced secure data management solutions. In the attempt of enabling these pilots, EUDAT has investigated the possibility to using state-of-art solutions, currently offered on national settings, for wider use across european-wide communities. In particular two solutions have been identified, namely the TSD (University of Oslo, Sigma2, Norway), and ePouta (CSC, Finland). TSD is a remote access solution to allow single users and institutions to store and compute on sensitive research data. All the data are stored in a centralised storage solution secured by a strong firewall. The secure cloud ePouta is an Infrastructure-as-a-service solution: the users (in this case institutions) are allowed to administer their own virtualized infrastructure at the Pouta cloud obtained as an extension of their local compute and storage resources. TSD and ePouta are operated by EUDAT partners and the services are also included in the upcoming EOSC-hub portfolio. TSD and ePouta represent complementary resources and thus they offer wide potential for integration to the European e-Infrastructures through the EUDAT and EOSC service portfolios. With the idea of enhancing the interoperability between secure services, work towards connecting TSD and ePouta with a secure connection is underway, providing a viable scenario for cross-border use of such infrastructures. Furthermore, the on-going integration of TSD with B2SHARE and B2ACCESS will enable publishing of metadata about sensitive data that is stored in the secure system, thus opening up the possibility of making sensitive data discoverable through anonymised metadata. Similar design will could later be used also for the integration between ePouta and B2SHARE & B2ACCESS.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

16 | Sensitive Data Group

Author: Wolfgang Kuchinke

Affiliation: Heinrich-Heine University Duesseldorf

Abstract:
Sensitive Data is personal data about a Data Subject`s racial origin, political opinions, religious beliefs and physical or mental health, including health and genetic data. Sensitive Data is subject to special protection by EU law. The “EUDAT Groups for Sensitive Data” addresses this challenge for the Open Science community and evaluates different solutions to enable the inclusion of sensitive data in the Data Fabric of advanced research processes. Our aim is that sensitive data should be accessed and analysed together with open data. For example, health data should be linked and analysed with climate and environmental data to finally improve the wellbeing of humans. Such an approach will foster all form of new health research, including health outcomes and comparative effectiveness research. To enable the promises of cloud computing, the European Open Science Cloud (EOSC) must therefore fully consider the inclusion of sensitive data, especially of health data, in its concepts for sharing and stewardship of research data.

The “EUDAT Group for Sensitive Data” discusses challenges for sensitive data, like certified authentication, integration of cloaking systems, authorisation by trusted Identity Providers. One solution is to connect B2SHARE to secure areas for data processing. In addition, we are evaluating the use of TSD (Tjenester for Sensitive Data), a system to collect, store and analyse sensitive data in a secure environment, for clinical study data and we monitor the impact of GDPR for the research community. In this context, the group organises workshops for exchange with infrastructures, especially with EUDAT, EGI, GÉANT and ELIXIR, and with different research communities (e.g. social and linguistic research, medical and environmental research) to match the possibilities of advanced technical solutions with the requirements and needs of researchers dealing with sensitive data challenges. Our concepts and results will be summarised in a “Memorandum for sensitive data use”.

17 | Metadata interoperability challenges in (distributed) research data infrastructures

Author: Timo Borst

Affiliation: ZBW – Leibniz Information Center for Economics, Kiel, Germany

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Research data (RD) management is on the rise, and so are initiatives that focus primarily on datasets as a working artifact. Standards and practices for describing RD across domains are still evolving. RD infrastructure initiatives with different scopes and consortia are already making their mark.
GeRDI – the Generic RD Infrastructure – is a project that aims to develop an infrastructure that will make long tail RD FAIR [1] and support interdisciplinary research. Due to diversity of involved communities in GeRDI, we believe that metadata interoperability between heterogeneous RD is a key factor. For a long tail RD infrastructure, another challenge is to balance both breadth and depth of RD so as to provide an interesting prospect for researchers [3].
Although with a focus on long tail of research data, GeRDI tackles some of the challenges that EOSC faces and eventual adjustments of our approach could potentially be generalized/scaled up for EOSC's needs:
- Standards and best practices required for EOSC could benefit from GeRDI,
- Interoperability between datasets,
- Diversity of research disciplines, an important recommendation for EOSC.
The main deliverable of our approach is the GeRDI metadata schema, which contains a "core" part to support the generic aspects, and the (conceptual) "Metadata bowl", to support the discipline-specific (metadata) requirements. For the former we rely on a generic metadata standard – DataCite metadata schema [4]; due to GeRDI operations requirements, we introduced a small metadata extension ("Functional" Metadata Extension), as shown in Figure 1; the latter constitutes of domain-specific metadata elements. The "Metadata bowl" can further be refined via categorizing metadata similar by certain condition, such as user preference, RD domain, etc., in order to better support information retrieval in GeRDI.
The two GeRDI schema constituents also provide different access levels: a coarse-grained one based on the schema "core", and a finer-grained one based on the metadata contained in "Metadata bowl". We expect this design to support a set of key services across disciplines (despite metadata diversity of RD), as well as a more fine grained access to the discipline-specific aspects of the RD. For example, a discovery service could enable search functionality to users based on schema "core" elements, such as author, title, or publication year, to name a few; and recommendation functionality based on relevant discipline-specific metadata elements in the "MD bowl" to enrich the discovery process of users.
We foresee several benefits from our approach with GeRDI, especially for:
o Metadata mapping and indexing: our approach especially considers these two operations in the light of RD metadata from different research disciplines;
o Reusing GeRDI schema: other projects could adapt to their needs by choosing a different balance or separation between the "core" metadata standard(s) and required "functional" extension.
GeRDI schema aims to make use of all harvested metadata elements, and not limit its services on a single, minimal metadata schema. We believe this design is capable of addressing RD Interdisciplinarity.

[1] https://www.force11.org/group/fairgroup/fairprinciples;
[2] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud;
[3] Long tail of data, e-IRG Task Force Report 2016
[4] https://schema.datacite.org/meta/kernel-4.0/

18 | openBIS ELN-LIMS: a comprehensive data management solution for the Big Data era

Author: Henry Luetcke
Affiliation: ETH Zurich

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Academic research labs nowadays produce ever larger amounts of data that have to be stored, interpreted and shared. Furthermore, new guidelines and policies regarding data management and, in particular sharing, are regularly issued by funding agencies, journals or academic institutions. As a result, successful research data management has become both a key skill and a major challenge for many scientists. The Scientific IT Services (SIS) group of ETH Zurich aims to support researchers at ETH Zurich and beyond with successful research data management. To this end, SIS has developed openBIS, an open source framework for creation of user-friendly, scalable and powerful information systems for data and metadata acquired in experimental research groups [1]. openBIS is highly flexible, versatile and extendable. It has been continuously developed and supported for the last 10 years and is currently used by a number of Swiss and European research groups, consortia and private companies for data management, including several Big Data projects (up to 100s of TB). Recently, openBIS has been extended with a combined Electronic Lab Notebook (ELN) and Laboratory Information Management system (LIMS) [2]. The ELN-LIMS provides a structured and easy-to-use interface for traceable storage and maintenance of information about experimental data, materials and methods in one central location. Further features of openBIS include the possibility to access managed data for analysis with the popular analysis frameworks Jupyter [3] and Matlab as well as the combination with workflow managers for large-scale data analytics on high-performance computing infrastructures. Moreover, SIS has also developed a tool that allows openBIS to track data stored on external, possibly remote, file systems where researchers retain full flexibility to structure and manipulate the data as they wish. In this scenario, openBIS has the capability to keep track of data from large-scale data analysis and simulations that require entire compute clusters or even supercomputers. Finally, openBIS features an open Java API which will allow it to be integrated into new federated infrastructures such as the European Open Science Cloud. The openBIS software, a user guide and a demo instance are available at https://openbis.elnlims.ch .

[1] Bauch et al. (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics 12:468
[2] Barillari et al. (2016) openBIS ELN-LIMS: an open-source database for academic laboratories. Bioinformatics 32:638
[3] Perez & Granger (2007) IPython: A System for Interactive Scientific Computing. Comput. Sci. Eng. 9:21

19 | Towards a pan-European Infrastructure of Scientific Collections

Author: Patricia Mergen
Affiliation: Botanic Garden Meise/ Royal Museum for Central Africa

Abstract:
DiSSCo (Distributed System of Scientific Collections) is a pan-European Research Infrastructure initiative of 21 European countries and 113 institutions. Its vision is to position European natural science collections at the centre of data-driven scientific excellence and innovation in environmental research, climate change, food security, health and bio-economy. The mission of DiSSCo is to mobilise, unify and deliver bio- and geo-diversity information at the scale, form and precision required by scientific communities ; evolving from the currently existing mature scientific community gathered under CETAF to a joint distributed research infrastructure, transforming a fragmented landscape into a coherent and responsive research infrastructure (www.DiSSCo.eu). Digitisation is giving access to millions of natural history related objects, and archives and associated information and this accessibility will enable novel research interpretations and applications in a wide range of domains, such as virtual reality, serious gaming or artificial intelligence. The actual implementation requires high quality digital curation, data mobilisation and data usage via e-infrastructure services applied to Big Data, such as cloud computing or High Performance Computing (HPC).

20 | EUDAT as a safe backup solution

Author: MJosef Misutka
Affiliation: LINDAT/CLARIN

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Digital repositories in the CLARIN project contain many unique and important data submissions. These can range from small ones having couple of megabytes to big ones having hundreds of gigabytes. Each digital repository has to meet the requirements of an external certification authority (e.g., CoreTrustSeal formerly known as Data Seal of Approval) that includes basic redundant storage of data. In order to help the digital repositories, a frontend platform for easy backups has been created by CLARIN that integrates EUDAT infrastructure and offers a safe long term preservation solution. Using EUDAT's HTTP-API abstraction layer, we were able to achieve this goal without needing to know the complex technology stack hidden behind it. The backup frontend consists of a two parts. A node.js application offering a very simple set of API calls decoupling it from the backup implementations. The backup frontend can be either deployed centrally or at individual centres. The second part is a very thin integration part available for the most common digital repository software used (DSpace and Fedora Commons). Nevertheless, we will also present how we dealt with some of the limits introduced by the HTTP-API e.g., upload size.

21 | Array Databases for Research Communities

Author: Christian Pagé

Affiliation: CERFACS

Abstract:
An Array Databases EUDAT Working Group is working together to address common issues and opportunities for Research Communities. This communities are dealing with multidimensional and spatial datasets. Following a workshop that was help earlier in May 2017 and organized by EUDAT, the poster will present real Use Cases that have been designed and experimented in research communities. Benefits with respect to traditional methods, limitations, challenges encountered and lessons learned will be shown.

22 | EUDAT GEF/C4I: Pushing Climate Post-Processing to the Cloud

Author: Asela Rajapakse

Affiliation: MPI-M

Abstract:
This poster presents the ongoing integration of the IS-ENES Climate4Impact platform which retrieves climate data for calculation (e.g. climate indices) and visualization from Earth Science Grid Federation data nodes using the Compute Working Team API in a Web Processing Service process with the EUDAT workflow service, the Generic Execution Framework (GEF). The aim is to perform climate calculations on the European Grid Infrastructure Federated Cloud by deploying the necessary software in a containerized version from the Climate4Impact platform through the GEF to a virtual machine on the Federated Cloud.

23 | DIGITAL PRESERVATION CHALLENGES AND METHODOLOGY USING CLOUD RESOURCES - THE OPENCOASTS CASE STUDY

Author: João Rogeiro

Affiliation: LNEC

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
In the era of exponential data growth, digital preservation and curation is of paramount importance to support science and engineering processes, ensuring the persistence of the scientific knowledge produced. Scientific research and development relies increasingly on software, large data sets and a workflow of different algorithms, which raise several challenges that could be partially addressed with cloud technologies. On the one hand, the large amount of available computing resources in cloud environments is bringing new research opportunities, including fast collaboration and transmission of knowledge at a fraction of the cost. On the other hand, inadequate practices of software management and control can easily lead to large costs in several fronts: incompatibilities between software versions, obsolescence, irreproducibility of results, and the inability to compare and contrast, in a meaningful way, different computational solutions to the same problem. This publication intends to address the combination of cloud computing with digital preservation strategies to contribute to the long-term persistence of scientific and engineering processes.

Within the Portuguese National Infrastructure for Distributed Computing (INCD) project, a new service - named OPENCoastS.pt - is being built to provide on-demand forecasts of the circulation in the Portuguese coast. This infrastructure provides advanced computing and data services to the academic and scientific community (covering all scientific domains). Its core services include HPC (high performance computing), cloud computing and HTC (high throughput computing), complemented with support, consulting and training activities. Its purpose is to support the participation of the Portuguese scientific community in national and international R&D projects. The INCD is partnered with several European initiatives, including EGI and IBERGRID.

We present a methodology for digital preservation in the INCD platform, highlighting novel aspects. The cloud/digital preservation case study OPENCoastS.pt uses a computational model based on SCHISM (Semi-implicit Cross-scale Hydroscience Integrated System Model). This an open-source community-supported modeling system for simulation of 3D baroclinic circulation across creek-lake-river-estuary-shelf-ocean scales, based on unstructured grids, using a semi-implicit finite-element/finite-volume method with Eulerian-Lagrangian algorithm to solve the Navier-Stokes equations. In particular we emphasize the importance of maintaining compatibility between different versions of the model amidst changes of input formats and different model parameters, a problem to be tackled in the use case OPENCoastS.pt and its European counterpart, named OPENCoastS, to be implemented in the scope of the EOSC-Hub H2020 project. We focus on 3 main scenarios: a) reproducing previous computations for trust, b) running previous models with new or updated data, c) running new models with old data. This enables the user to navigate between different versions of the model, comparing and assessing its quality, ensuring the capability to preserve not only the data, but its model version and execution context. This requires attention to several aspects of digital preservation.

The proposal is a promising solution to the problems of digital preservation in scientific and engineering contexts, providing the capability to reproduce research.

24 | DOMAIN-SPECIFIC METADATA IN B2SHARE RECORDS

Author: Joana Rodrigues
Affiliation: INESC TEC/FEUP

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
The Dendro platform is an open-source data storage and description platform designed to help users describe their data files and is fully built on Linked Open Data and ontologies. Moreover, it is a collaborative platform, capturing metadata as soon as researchers start to produce their data. Whenever users want to publish a dataset, they can export the corresponding project or folder to a repository such as CKAN, Zenodo, DSpace, or EUDAT's B2SHARE. Hence, Dendro complements data repositories in the early stage of the research data workflow.
Dendro manages metadata with flexible models. With EUDATLink, the gateway between Dendro and EUDAT built in the UPorto Pilot, data are deposited and Dublin Core descriptive metadata are transformed into B2SHARE metadata. Domain-specific metadata, however, is only present in the associated metadata files, which are not used to index or find the datasets. With a proper infrastructure and tools in place, researchers will be able to share data and exploit it to the full to derive new knowledge.
We report here a set of case studies from the long tail of science. The groups participating in these cases interacted with us at 3 points in time: 1) by discussing their metadata models with a data curator and defining a metadata model; 2) by describing datasets with the metadata model they adopted; 3) by comparing the full metadata records with the ones displayed in B2SHARE.
Each case study resulted from the interaction with a research group from the University of Porto that has described their data using the Dendro platform, using domain-specific metadata. The data packages, with both the datasets and the metadata, were then transferred from Dendro to B2SHARE. Together, the selected cases capture data description requirements from a diversity of research domains and from projects with a different scope, including observational, experimental and computational data.
The research data deposited in B2SHARE include the outcome of social science related studies, such as partial data from a longitudinal psychology project, and data from two projects in innovation management and social media communication analysis. The experimental and computational data were provided by groups in electrical and computer engineering, computational engineering, and materials engineering domains. The results show that researchers can provide metadata elements that best fit the context of their data, thus improving the access and interpretation of the datasets by B2SHARE users. In our case studies researchers have selected metadata elements from the Data Documentation Initiative, Dublin Core and the Friend of a Friend vocabularies, but also from domain-specific metadata models developed at the University of Porto, previously represented in the Dendro platform.
The results, comparing the B2SHARE metadata with those in the full record captured with Dendro, suggest that the B2SHARE metadata representation may need to evolve to more flexible models. Further work is necessary to expand this study to more domains and contribute to the goal of flexible metadata in B2SHARE.

25 | Data Federation Hub: Building a Local Hub for Research Data Services

Author: Anna Salzano

Affiliation: DFH/ CIT, University of Groningen (NL)

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
In order to empower Open science and FAIR data in the Netherlands, researchers need a strong research infrastructure, with local direct access to FAIR data, IT solutions, data stewardship, and ethical and legal support. Building on existing excellence, the University of Groningen and the University Medical Center Groningen, in connection with (inter)national infrastructures, joined efforts to set up such local, integrated research support platform: the Data Federation Hub (DFH). Following a hub-and-spokes-model of interconnected services, the DFH comprehends sub-chapters for specific data-domain (e.g. Human data, Astronomy, e-Humanities etc.), and supports the whole research data lifecycle: from Preparation, Collection, Processing, Analysis, Preservation, Access, and Re-Use. DFH enables researchers to collaborate, share, harmonize, and profit from the resources available locally and (inter)nationally.
The DFH is developing an online platform where local researchers can easily get access to relevant tools and services for data management. Via researcher-relevant browsing features, user-oriented filtering, and advanced search capability, researchers at UG/UMCG will be able to access the most relevant tools and services (e.g. collaborative Virtual Research Environments, secure data storage and archiving solutions, and ad-hoc Data Management tools) but also policies, practices and expertise. A team of data consultants is made available to translate the often complex multidisciplinary questions from research problems to solutions. This platform also aims at promoting collaboration and communication among researchers, service providers, and experts, with the aim to share best practices and further innovation regionally.

The Tools and services made available via the Data Federation Hub are aimed at allowing for FAIR and open access, while also ensuing security and efficient protection of the privacy of participants when dealing with human subject data. When dealing with Big data, tools and expertise is made available to store, process and analyse excessively large datasets. The DFH therefore facilitates not only access to tools and IT solutions, but direct, local connection with experts and specialists on data, i.e. data stewards, data scientists, legal experts in the field of data, privacy and IP, and security experts.

26 | Promoting Open Science through new The Bridge of Knowledge platform at Gdansk University of Technology

Author: Magdalena Szuflita-Żurawska, Anna Wałek

Affiliation: Gdansk University of Technology

Abstract:
The process of scholarly communications and adopting Open Science principles is changing dramatically. It has matured significantly by using a proper infrastructure for sharing and dissemination of scientific knowledge. Nowadays, creating awareness among academic community and offer solutions to specific issues are the most important goals for the Library. We would like to present challenges and strategies to promote scientific output of Gdansk University of Technology (GUT), Poland through new established platform “MOST Wiedzy”. MOST Wiedzy in Polish means The Bridge of Knowledge and stands as acronym of Multidisciplinary Open System Transferring Knowledge. It should be stressed that this project is co-financed by the European Regional Development Fund within the Optional Programme Digital Poland for the years 2014 – 2020.
MOST Wiedzy project complements other features under the umbrella of Open Science platform. The platform is a solution that supports not only scholarly communication but try to facilitate the cooperation between university and business industry. The main goals include for example: integrating data from different databases and providers as well as establishing open repository.
MOST Wiedzy is a modern platform that allows to present scientific profiles and research conducted at GUT. Through its website, users can search for publications, teaching activities or research equipment. MOST Wiedzy offers repository support from team of experienced librarians that not only upload the open access manuscripts but in addition, steer publications through different phases such as editing or promoting.
In our poster, we would like to demonstrate how MOST Wiedzy project helps researcher to share their scientific results, gain recognition and promote their achievements.

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.

27 | Preparing for the EOSC: Data-related training and community building for researchers, developers and trainers

Author: Celia van Gelder

Affiliation: DTL Dutch Techcentre for Life Sciences / ELIXIR-Netherlands

Abstract:
It is evident that in the EOSC era there is a demand for a substantial number of data experts, on different levels, and with new skill sets, which are currently being identified and defined. Researchers, programmers and educators need the training to acquire those skills. Also, the frontrunners in the field, e.g. the data stewards that start to populate these new professions, do need to share experiences and build this new community together. To tackle these challenges, DTL/ELIXIR-NL associated professionals bring together their expertise in a national network that crosses disciplines, application domains and European research infrastructures (ESFRIs and e-infrastructures).

Examples of our current activities are:

- FAIR data training, Bring Your Own Data Workshops (BYODs) and training in research Data Management and Data Stewardship
- Setting up national training strategies for Galaxy and Software and Data Carpentry (SWC/DC)
- Training in bioinformatics, systems biology, NGS, metabolomics and e-infrastructure access
- Train the trainer activities to build training capacity
- The FAIR Data Interest Group and the Data Stewards Interest Group. Interest Groups are the perfect way to learn about the views of others, to explore solutions, and to evaluate whether a collaborative effort is needed and justified.

28 | Uptake of EUDAT services within the CLARIN infrastructure

Author: Dieter Van Uytvanck

Affiliation: CLARIN ERIC

Abstract:
With the CLARIN uptake plan CLARIN ERIC supports CLARIN centres to increase uptake of the EUDAT services, such as B2SAFE, B2STAGE and B2DROP, by liaising between the CLARIN and EUDAT stakeholders and providing technical support where needed. This poster aims to provide an overview, per EUDAT service, of the progress made on the integration of the EUDAT services into the various CLARIN centres.

29 | THE ICOS Carbon Portal Service Portfolio

Author: Alex Vermeulen

Affiliation: ICOS ERIC

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
The ICOS Carbon Portal (CP) is the one stop shop data platform of ICOS.
ICOS consists of a distributed network of observational stations in three domains: atmosphere, ecosystem and ocean. Each domain data is processed in (again) distributed thematic centers that perform the quality control, calibration and condensing of the data to final data products.. The Carbon Portal takes care of curation of all raw data and processed observational data products and publishes them to the users and other portals and portals of portals.
The Carbon Portal has also the role to coordinate, facilitate and ensure production of elaborated products based on ICOS data in direct collaboration with the modelling community. Global as well as regional flux and emission datasets are collected and will be analysed and displayed at the CP.
The Carbon Portal has been developed from the ground up as a PID central, open linked data archive based on ontology and semantic web technology. Central in the Carbon Portal is a data ingestion engine, optimised for machine to machine interface, that ingests all agreed input data, mints PIDs (DOI in case of final data products), stores the data, streams it to the trusted repository (B2SAFE) and connects the PID of the data to the metadata system. The metadata system is a versioned triple store connected to an open SPARQL endpoint for data discovery and retrieval. All PIDs resolve to dynamically generated landing pages that present the relevant metadata for the connected data object. The system also supports data usage tracking and automatic generation of citations for the data based on the ontology.
In this poster we will present the services provided by the Cabron Portal and how these are connected to the system design. The setup of the Carbon Portal ingestion engine and connected meta data triple store with connected services that rely on the CDI/EUDAT and EGI services is very generic, open source and modular and is proposed as an interesting building block for constructing scalable, interoperable data portals.

30 | A Bridge from EUDAT's B2DROP cloud service to CLARIN's Language Resource Switchboard

Author: Claus Zinn

Affiliation: University of Tuebingen

Abstract:
The CLARIN Language Resource Switchboard (LRS) aims at bridging the gap between language-related resources and tools that can deal with these resources in one way or another. For a given resource, the LRS identifies all tools that can process the resource; users can then select and invoke the tool of their choosing. By invoking the tool, all relevant information about the resource is passed onto the tool, and the tool opens with most information gathered by the switchboard. This makes it easy for users to identify the right tools for their resource, but also to use the chosen tool in the most effective way possible.

The EUDAT Collaborative Data Infrastructure aims at providing services that seek to address the full life-cycle of research data. EUDAT's services include, among others, B2DROP (sync and exchange of research data), B2SHARE (store and share research data), B2FIND (find research data), and B2HANDLE (register your research data). B2DROP is directed at scientists to store and exchange data easily and to facilitate data synchronisation between cloud storage and desktop computers. EUDAT services are designed, built and implemented based on user community requirements. The CLARIN consortium contributes to EUDAT as one of the main communities in the Social Sciences and Humanities.

In this poster, we describe the use of B2DROP in the CLARIN Language Resource Switchboard. In the main use case, we anticipate an individual researcher or a small team of researchers to use B2DROP as cloud storage for language-related resources. The researcher(s) will want to work with and analyse the resources using community-specific tools of the CLARIN tool space. From the B2DROP user interface, the researcher(s) will want to easily transfer a given resource to the LRS, which in turn suggests tools to process the resource. In a second use case, we describe the use of B2DROP as a technical vehicle for intermediate cloud storage, supporting a crucial aspect of the LRS' back-end implementation. Both use cases have been implemented. Locally-configured B2DROP instances have been set-up; the proof-of-concept implementations are available for testing.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

31 | Handling Big Data and Sensitive Data Combining EUDAT's Generic Execution Framework with WebLicht

Author: Claus Zinn

Affiliation: University of Tuebingen

Abstract:
Our work addresses two challenges that affect the applicability of workflows for some data sets. First, restrictive property rights may forbid research data to leave their home institution, and therefore, data transferal to other institutions for processing is not allowed. The second issue concerns the size of the data, with big data often causing prohibitive overhead once it is necessary to send such data back and forth to the various tools of a scientific tool pipeline. In both cases, it is desirable to bring the workflow engine to the data, rather than having the data travel to the tools.
The EUDAT project is currently developing the Generic Execution Framework (GEF). The GEF aims at providing a framework that allows the execution of scientific workflows in a computing environment close to the data. In this paper, we describe how WebLicht needs to be adapted to render its services compatible with the GEF.
WebLicht is a workflow engine giving users a web-based access to over fifty tools for analysing digital texts. Its pipelining engine offers predefined workflows (``Easy Mode'') and supports users in configuring their own (``Advanced Mode''). With WebLicht, users can analyse texts at different levels such as morphology analysis, tokenization, part-of-speech tagging, lemmatization, dependency parsing, constituent parsing, co-occurrence analysis and word frequency analysis, supporting mainly German, English, and Dutch. Note that WebLicht does not implement any of the tools itself but mediates their use via pre-defined as well as user-configurable process pipelines. These workflows schedule the succession of tools so that one tool is called after another to achieve a given task, say, named entity recognition. WebLicht is a good step forward in increasing (web-based) tool access and usability as its TCF format mediates between the various input and output formats the tools require, and calls the tools (hosted on many different servers located nation and world-wide) without any user engagement. WebLicht has now been used for many years in the linguistics community, and it is the workflow engine of choice for many national and European researchers in the CLARIN context. WebLicht is actively maintained, and profits from regular tool updates and new tool integrations.

In this poster, we show how we (i) adapted the WebLicht orchestrator to invoke tools by posting them URLs pointing to the data, rather than the actual data; (ii) built wrappers around tools to accept URLs rather than actual data; (iii) set-up a local GEF environment with a number of language-related processing services; and (iv) devised BridgIT, a liasion device to facilitate the communication between WebLicht and a GEF environment. These steps resulted in a proof-of-concept implementation that demonstrates how WebLicht-orchestrated tool chains can run in a GEF environment close to the data.

Our work informed the EUDAT-GEF development team to ensure that user requirements stemming from the WebLicht use case lead to GEF feature requests that will find their way into the official GEF specification and implementation.

32 | Materials Cloud: A Platform for Open Materials Science

Author: Snehal Kumbhar

Affiliation: EPLF, Lausanne

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
Materials Cloud (www.materialscloud.org) is a web platform designed to enable seamless sharing of resources in computational materials science, including educational material, interactive tools and virtual hardware to run simulations, up to publishing results in a FAIR-compliant format [1]. Materials Cloud is powered by AiiDA [2], a python framework to manage materials science calculations, automatically storing the full provenance of data and calculations.
By share scientific data on Materials Cloud, not only the results of calculations but every step along the way is made available and is fully downloadable, both as individual files or as a whole database, so that research results can be seamlessly reused. Moreover, the web interface makes it easy to browse and query for calculations and data, as well as it provides a Jupyter-based interface to run simulations in the cloud. Finally, DOIs are assigned to published results to make them persistent and citable.
The Materials Cloud web interface is being developed using modern web technologies. On the server side there is the AiiDA API exposed via a REST interface, while the client uses libraries like AngularJS, D3.js, Bootstrap, JSmol, JQuery. CSCS is developing a stack of federated services for authentication and authorization (KeyStone), object storage (Swift) and web services (OpenStack), to be extended to CINECA (Italy) and Jülich (Germany) to build a decentralized cloud.

[1] M. D. Wilkinson et al., Scientific Data 3, 160018 (2016)
[2] G. Pizzi et al., Comp. Mat. Sci. 111, 218 (2016) - www.aiida.net

33 | Support to scientific research on seasonal-to-decadal climate and air quality modelling

Author: Francesco Benincasa

Affiliation: BSC

Abstract:
The Sand and Dust Storms Warning Advisory and Assessment System (SDS-WAS) is an international framework under the umbrella of the World Meteorological Organization (WMO). Its mission is to enhance the ability of countries to deliver timely and quality sand and dust storm forecasts, observations, information and knowledge to users through an international partnership of research and operational communities. In this EUDAT pilot we use the combination of B2SAFE federated storage with B2STAGE/http-api to make available to community users a huge amount of dust forecast datasets coming from a variety of international partners (or executed on site) through a web interface and an API in a transparent way.

34 | EUDAT B2FIND – Making Research Data FAIR

Author: Claudia Martens, Heinrich Widmann

Affiliation: DKRZ, Hamburg Germany

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Abstract:
The metadata service B2FIND plays a central role within the pan-European collaborative research data infrastructure EUDAT-CDI by providing a simple and user-friendly discovery portal to find and access research data collections stored in EUDAT data centers or in other repositories. Therefore metadata collected from heterogeneous sources are stored in a comprehensive joint metadata catalogue and made searchable via an open data portal. Furthermore B2FIND provides transparent access to the scientific data objects through the given references and identifiers in the metadata – thus supporting the first two pillows of FAIR data principles.

The implemented metadata ingestion workflow consists of three steps. First the metadata records - provided either by various research communities or via other EUDAT services - are harvested. Afterwards the raw metadata records are converted and mapped to unified key-value dictionaries as specified by the B2FIND schema. The semantic mapping of the non-uniform, community specific metadata to homogenous structured datasets is hereby the most subtle and challenging task. To assure and improve metadata quality this mapping process is accompanied by
• iterative and intense exchange with community representatives,
• usage of controlled vocabularies and community specific ontologies and
• formal and semantic mapping and validation.
Finally the mapped and checked records are uploaded as datasets to the catalogue which is based on CKAN, an open source data portal software that provides a rich RESTful JSON API and uses SOLR for indexing.
The homogenization of community specific data models and vocabularies enables not only a unique presentation of these datasets as tables of field-value pairs but also an interdisciplinary and cross-community search with geospatial and temporal search functionalities. Results from a free text search may be narrowed by using the facets. B2FIND offers support for new communities interested in publishing their data within EUDAT and EOSC.

35 | Herbadrop Data Pilot: extracting knowledge from distributed herbaria accross Europe

Author: Pascal Dugénie

Affiliation: Centre Informatique National de l'Enseignelment Supérieur (CINES), Montpellier, France

Abstract:
Herbadrop is one of the EUDAT data pilot that enables long-term preservation of herbarium specimen images and aims to extract information from these images by using Optical Character Recognition (OCR) analysis. Making the specimen images and data available on-line from different institutes allows cross domain research and data analysis for botanists and researchers with diverse interests. Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. High resolution images of these specimens require substantial bandwidth and disk space. New methods of extracting information from the specimen labels have been developed using OCR but using this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts. Much of the information is only available only using handwritten text recognition or botanical pattern recognition which is less mature technology than OCR.

EUDAT B2SAFE service is used in the first step of the ingestion process in accordance with the centralized persistent identifiers (PID) management. Different processes in the workflow enable to develop approaches for building knowledge using emerging methods and big data technologies together with archival practices.

This poster presents the workflow and the results publised during the year 2017. It shows how technologcal trends may offer some new research potential in the domain of computational archival science in particular appraising the challenges of producing quality, meaning, knowledge and value from quantity, tracing data and analytic provenance across complex big data platforms and knowledge production ecosystems.

Read more

EUDAT receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654065.